Table 3. Summary of major approaches for detection of machine generated text

Approach summary	Base model	Releated research	Stat. features	NLM features	Evaluated Against			
					GPT-2	GPT-3	Grover	Other Datasets/Models
Algorithmic Detection	K-nearest-neighbor	Lavoie et al. 2010 [104]	✓					SCIgen
Statistical Features	SVM	Nguyen-Son et al. 2017 [128]	✓					Google Translate
TF-IDF Baseline	LR	Radford, Wu et al. 2019 [138] Solaiman et al. 2019 [163]	✓		✓			
Zero-shot GPT-2	GPT-2	Radford, Wu et al. 2019 [138] Zellers et al. 2019 [190] Solaiman et al. 2019 [163]		✓	✓			
Zero-shot Grover	Grover	Zellers et al. 2019 [190] Solaiman et al. 2019 [163]		✓	✓		✓	
GLTR	BERT, GPT-2	Gehrmann et al. 2019 [66] Ippolito et al. 2019 [79]		✓	✓			
RoBERTa fine-tuning	RoBERTa	Solaiman et al. 2019 [163]		✓	✓			
Energy Based Models	BiLSTM, GPT, RoBERTa	Bakhtin et al. 2019 [12]		✓	✓			
Feature Ensemble	LR, SVM, RF, NN	Fröhling et al. 2021 [60]	✓		✓	✓	✓	
Twitter-specific RoBERTA fine-tuning	RoBERTa	Fagni et al. 2021 [57] Tourille et al. 2022 [172]		✓	✓			TweepFake (incl. RNN/LSTM/Markov)
Human-Bot Interaction Feat. Ensemble	BERT, LR	Bhatt and Rios, 2021 [22]	✓	✓				ConvAI2, WOCHAT, DailyDialog
Neural-Stat. Ensemble	RoBERTa, SVM	Crothers et al. 2022 [40]	✓	✓	✓	✓		
Explainable classifiers	RF, XGBoost	Kowalczyk et al. 2022 [97]	✓		✓			
Disinformation-specific RoBERTA fine-tuning	RoBERTa	Stiff et al. 2022 [166]		✓	✓	✓	✓	TweepFake, XLM, PPLM, GeDi

of the study, however, these raters had consistently worse accuracy than automatic classifiers for all sampling methods (random, top-k, and nucleus) and excerpt lengths.

Further demonstrating the advantage of providing specialized training to human reviewers, the Scarecrow framework specifically identifies 10 categories of common errors made in GPT-3 generative text, and trains human evaluators to annotate these errors [51]. Human annotations of such errors were found to generally be of higher precision than a corresponding algorithm trained on such annotations, but had higher F_1 scores in only half of the categories.

Based on these findings, we can better inform defenses against threat models. For example, in the social media domain, it is possible that if a social media company hired a specialist human moderator and provided them